Distributionnally Robust Optimization & Statistical Learning

Franck Iutzeler

TSE – Oct 19th, 2023

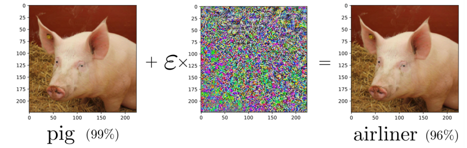
INSTITUT de MATHEMATIQUES de TOULOUSE

UNIVERSITÉ TOULOUSE III PAUL SABATIER

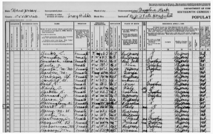
- Mathematical modelling
  - The **cost** f_x of a decision **parametrized** by $x \in \mathcal{X}$
  - depends on an **uncertain variable** $\xi \in \Xi$
- Why do we want **robustness** in practical applications?

Difficult-to-predict environments

Attacks against complex models



Biased, outdated, insufficient data



In phase with regulations

◇ Ben-Tal, Ghaoui, Nemirovski. *Robust optimization.* Princeton university press, 2009.
◇ Kolter, Madry. *Adversarial robustness - theory and practice.* NeurIPS tutorial https://adversarial-ml-tutorial.org/, 2018.

Decision under uncertainty

- Mathematical modelling
  - The **cost** f_x of a decision **parametrized** by $x \in \mathcal{X}$
  - depends on an **uncertain variable** $\xi \in \Xi$
- Why do we want **robustness** in **statistical learning**?
  - cost = model + loss f_x on data point ξ ex. least squares $f_x(\xi = (a, b)) = (\langle x, a \rangle - b)^2$
  - the uncertainty variable's **distribution** is known through samples $\xi_1, .., \xi_N$
  - Robustness is desirable for
    - Generalization guarantees on the true distribution of the samples
    - Distribution shifts between training and application

Popular approaches

- The *uncertain variable* ξ lives in some **uncertainty set** U

$$\min_{x \in \mathcal{X}} \sup_{\xi \in U} f_x(\xi) \qquad \text{(Worst-case robustness)}$$

  - U may be difficult to design
  - pessimistic decisions (unlikely values of ξ)

- The *uncertain variable* ξ is known though its **empirical distribution** $\hat{\mathbf{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\xi_i}$

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\xi \sim \hat{\mathbf{P}}_N}[f_x(\xi)] \qquad \text{(Sample Average Approximation)}$$

  - also called Empirical Risk Minimization in machine learning
  - the empirical distribution $\hat{\mathbf{P}}_N$ may not be close to the true distribution of ξ in the target application

    too few samples, biased collection, distribution shifts

- Ben-Tal and Nemirovski. *Robust convex optimization.* Mathematics of operations research, 1998.
- Shapiro, Dentcheva, and Ruszczynski. *Lectures on stochastic programming: modeling and theory.* SIAM, 2021.

Distributionally Robust Optimization

- The empirical distribution data provides **partial information** about the encountered **distribution** of ξ
  - The uncertain variable's **distribution** lives in a **neighborhood** $\mathcal{U}(\hat{\mathbf{P}}_N)$ of its empirical distribution

$$\min_{x \in \mathcal{X}} \sup_{\substack{\mathbf{Q} \in \mathcal{P}(\Xi) \\ \mathbf{Q} \in \mathcal{U}(\hat{\mathbf{P}}_N)}} \mathbb{E}_{\xi \sim \mathbf{Q}}[f_x(\xi)] \tag{DRO}$$

  - Inner sup taken over the set $\mathcal{P}(\Xi)$ of probability measures on Ξ infinite dimensional
  - For some $\mathcal{U}(\hat{\mathbf{P}}_N)$, parametric (Gaussian) or not (ϕ-divergences), this leads to finite-dimension min-max problems efficient stochastic optimization methods
  - Enforces **model robustness at training**

- Scarf. *A min-max solution of an inventory problem.* Studies in the mathematical theory of inventory and production, 1958.
- Rahimian and Mehrotra. *Distributionally robust optimization: A review.* arXiv 1908.05659, 2019.
- Delage and Ye. *Distributionally robust optimization under moment uncertainty with application to data-driven problems.* Op. Res., 2010.
- Namkoong and Duchi. *Stochastic gradient methods for distributionally robust optimization with f-divergences.* NeurIPS, 2016.

Wasserstein Distributionally Robust Optimization

- The uncertain variable's **distribution** lives in a **Wasserstein neighborhood** of its empirical distribution

$$\min_{x\in\mathcal{X}} \sup_{\substack{\mathbf{Q}\in\mathcal{P}(\Xi)\\ W_c(\hat{\mathbf{P}}_N,\mathbf{Q})\leq\rho}} \mathbb{E}_{\xi\sim\mathbf{Q}}[f_x(\xi)] \tag{WDRO}$$

  - For a cost function $c : \Xi\times\Xi \to \mathbb{R}_+$, the Wasserstein distance between $\hat{\mathbf{P}}_N$ and $\mathbf{Q}$ is defined as

$$W_c(\hat{\mathbf{P}}_N,\mathbf{Q}) = \inf\left\{\mathbb{E}_{(\xi,\zeta)\sim\pi}\left[c(\xi,\zeta)\right] : \pi\in\mathcal{P}(\Xi\times\Xi), \pi_1 = \hat{\mathbf{P}}_N, \pi_2 = \mathbf{Q}\right\},$$

  with π_1 (resp. π_2) the first (resp. second) marginal of the transport plan π.

  - **Natural metric** to compare empirical and absolutely continuous distributions contrary to the Kullback-Leibler divergence and strong generalization/concentration results

  - Inner sup stays infinite dimensional and the constraint is itself linked to an optimization problem

- Esfahani and Kuhn. *Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations.* Mathematical Programming, 2018.
- Kuhn, Esfahani, Nguyen, and Shafieezadeh-Abadeh. *Wasserstein distributionally robust optimization: Theory and applications in machine learning.* In Operations Research & Management Science in the Age of Analytics, 2019.
- Blanchet and Murthy. *Quantifying distributional model risk via optimal transport.* Mathematics of Operations Research, 2019.
- Gao and Kleywegt. *Distributionally robust stochastic optimization with Wasserstein distance.* Mathematics of Operations Research, 2022.

- WDRO is an appealing framework for distributional robustness but difficult to optimize
  - Understand precisely the behavior of WDRO solutions
  - Study its statistical guarantees
  - Provide computationally tractable formulations for a large class of problems

Outline

Wasserstein Distributionally Robust Optimization

- ◊ **Diving into the problem**

Diving into the problem

- Let us investigate the WDRO inner problem (we drop the $\min_x$ part)
  - Make explicit the Wasserstein constraint

$$\widehat{\mathcal{R}}_\rho(f_x) := \sup_{\substack{\mathbf{Q}\in\mathcal{P}(\Xi) \\ W_c(\hat{\mathbf{P}}_N,\mathbf{Q})\leq\rho}} \mathbb{E}_{\xi\sim\mathbf{Q}}[f_x(\xi)] \quad \textit{with} \quad W_c(\hat{\mathbf{P}}_N,\mathbf{Q}) = \inf_{\substack{\pi\in\mathcal{P}(\Xi\times\Xi) \\ \pi_1=\hat{\mathbf{P}}_N,\pi_2=\mathbf{Q}}} \mathbb{E}_{(\xi,\zeta)\sim\pi}\left[c(\xi,\zeta)\right]$$

Diving into the problem

- Let us investigate the WDRO inner problem (we drop the $\min_x$ part)
  - Make explicit the Wasserstein constraint $\mathbf{Q}$ disappears
  - Use the topological duality between signed measures and continuous functions on compact spaces

$$\widehat{\mathcal{R}}_\rho(f_x) := \sup_{\substack{\mathbf{Q}\in\mathcal{P}(\Xi) \\ W_c(\hat{\mathbf{P}}_N,\mathbf{Q})\leq\rho}} \mathbb{E}_{\xi\sim\mathbf{Q}}[f_x(\xi)] = \sup_{\substack{\pi\in\mathcal{P}(\Xi\times\Xi) \\ \pi_1=\hat{\mathbf{P}}_N,\ \mathbb{E}_{(\xi,\zeta)\sim\pi}[c(\xi,\zeta)]\leq\rho}} \mathbb{E}_{\zeta\sim\pi_2}[f_x(\zeta)]$$

- Let us investigate the WDRO inner problem (we drop the $\min_x$ part)
  - Make explicit the Wasserstein constraint $\mathbf{Q}$ disappears
  - Use the topological duality between signed measures and continuous functions on compact spaces
  - Denoting by $\langle \cdot, \cdot \rangle$ the corresponding duality pairing so that $\langle \pi, \varphi \rangle := \int_X \varphi(x) \,\mathrm{d}\,\pi(x)$ with $\tilde{f}_x : (\xi, \zeta) \mapsto f_x(\zeta)$ and c assumed to be continuous Riesz representation theorem

$$\widehat{\mathcal{R}}_\rho(f_x) := \sup_{\substack{\mathbf{Q} \in \mathcal{P}(\Xi) \\ W_c(\hat{\mathbf{P}}_N, \mathbf{Q}) \leq \rho}} \mathbb{E}_{\xi \sim \mathbf{Q}}[f_x(\xi)] = \sup_{\substack{\pi \in \mathcal{M}(\Xi \times \Xi) \\ \pi_1 = \hat{\mathbf{P}}_N,\ \langle \pi, c \rangle \leq \rho}} \langle \pi, \tilde{f}_x \rangle$$

- Let us investigate the WDRO inner problem (we drop the $\min_x$ part)
  - Make explicit the Wasserstein constraint **Q** disappears
  - Use the topological duality between signed measures and continuous functions on compact spaces
  - Denoting by $\langle \cdot, \cdot \rangle$ the corresponding duality pairing so that $\langle \pi, \varphi \rangle := \int_x \varphi(x) \, \mathrm{d}\, \pi(x)$ with $\tilde{f}_x : (\xi, \zeta) \mapsto f_x(\zeta)$ and c assumed to be continuous Riesz representation theorem

$$\widehat{\mathcal{R}}_\rho(f_x) := \sup_{\substack{\mathbf{Q} \in \mathcal{P}(\Xi) \\ W_c(\hat{\mathbf{P}}_N, \mathbf{Q}) \leq \rho}} \mathbb{E}_{\xi \sim \mathbf{Q}}[f_x(\xi)] = \sup_{\substack{\pi \in \mathcal{M}(\Xi \times \Xi) \\ \pi_1 = \hat{\mathbf{P}}_N,\ \langle \pi, c \rangle \leq \rho}} \langle \pi, \tilde{f}_x \rangle$$

- This is a **linear program** on measures
  - The solutions belong to the **border** of the constraint set
  - The optimal **worst-case distribution** $\pi_2^\star$ is supported on $N+1$ atoms LP with $N+1$ constraints

- Pinelis. *On the extreme points of moments sets.* Mathematical Methods of Operations Research, 2016.
- Yue, Kuhn, and Wiesemann. *On linear optimization over Wasserstein balls.* Mathematical Programming, 2021.

Dual problem

- **Duality** is at the core of modern WDRO
  - Lagrangian duality + Sup over (conditional) measure realized by a Dirac at the sup

$$\widehat{\mathcal{R}}_\rho(f_x) = \sup_{\substack{\pi \in \mathcal{P}(\Xi\times\Xi) \\ \pi_1 = \hat{\mathbf{P}}_N,\ \mathbb{E}_{(\xi,\zeta)\sim\pi}[c(\xi,\zeta)] \leq \rho}} \mathbb{E}_{\zeta\sim\pi_2}[f_x(\zeta)]$$

$$= \inf_{\lambda\geq 0} \lambda\rho + \mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\sup_{\zeta\in\Xi}\{f_x(\zeta) - \lambda c(\xi,\zeta)\}\right] \tag{Dual-WDRO}$$

- Main improvement: this is a finite-dimensional problem and λ is 1D!
  - **If** the sup is tractable, the Dual-WDRO problem is solvable! and thus WDRO, but that's a big if
  - The optimal **worst-case distribution** is supported on $N+1$ atoms taken in $\arg\max_{\zeta\in\Xi}\{f_x(\zeta) - \lambda^\star c(\xi_i,\zeta)\}$ for $i = 1, .., N$

- Esfahani and Kuhn. *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations.* Mathematical Programming, 2018.
- Zhao and Guan. *Data-driven risk-averse stochastic optimization with Wasserstein metric.* Operations Research Letters, 2018.
- Blanchet and Murthy. *Quantifying distributional model risk via optimal transport.* Mathematics of Operations Research, 2019.
- Gao and Kleywegt. *Distributionally robust stochastic optimization with Wasserstein distance.* Mathematics of Operations Research, 2022.

- Putting it all together, we have to solve

$$\min_{x\in\mathcal{X}} \overbrace{\sup_{\substack{\mathbf{Q}\in\mathcal{P}(\Xi)\\ W_c(\hat{\mathbf{P}}_N,\mathbf{Q})\leq\rho}} \mathbb{E}_{\xi\sim\mathbf{Q}}[f_x(\xi)]}^{\widehat{\mathcal{R}}_\rho(f_x)} = \min_{x\in\mathcal{X}} \inf_{\lambda\geq 0} \lambda\rho + \mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\sup_{\zeta\in\Xi}\{f_x(\zeta)-\lambda c(\xi,\zeta)\}\right]$$

- min/min problem easier than previous min-max but with an inner sup bottleneck
- Strong interplay between the loss f_x and the transport cost c
  - With x fixed, λ should be such that $\sup_{\zeta\in\Xi}\{f_x(\zeta)-\lambda c(\xi,\zeta)\} < +\infty$
  - If Ξ is bounded and the loss and cost are Lipschitz, this is ok
  - If Ξ is unbounded, $\frac{f_x(\zeta)}{c(\xi,\zeta)}$ should be uniformly bounded

Example I – the NewsVendor problem

- A NewsVendor has to decide how many papers he will buy for tomorrow
  - His buying price is $k = 5$ and his retail price is $u = 7$
  - He has a collection of sales data $\xi_1, .., \xi_N$
  - He wants to minimize its loss $f_x(\xi) = kx - u \min(x, \xi)$ by optimizing the number $x \in \mathbb{R}_+$ of newspaper bought, facing the uncertain demand of tomorrow $\xi \in \mathbb{R}_+$

- Taking a robust decision
  - Worst-case robustness leads to $x^\star_{WCR} = 0$ since $\xi = 0$ is possible
  - Sample Average Approximation leads to $x^\star_{SAA} > 0$ by minimizing the average loss over the past
  - What about WDRO?

Example I – the NewsVendor problem

- A NewsVendor has to decide how many papers he will buy for tomorrow
  - His buying price is $k = 5$ and his retail price is $u = 7$
  - He has a collection of sales data $\xi_1, .., \xi_N$ in $\mathbb{R}_+ = \Xi$
  - He wants to minimize its loss $f_x(\xi) = kx - u\min(x, \xi)$ by optimizing the number $x \in \mathbb{R}_+$ of newspaper bought, facing the uncertain demand of tomorrow $\xi \in \mathbb{R}_+$

$$\min_{x \geq 0} \inf_{\lambda \geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^{N} \sup_{\zeta \in \Xi} \left\{ kx - u\min(x, \zeta) - \lambda|\xi_i - \zeta| \right\}$$

- We can solve Dual-WDRO with $c(\xi, \zeta) = |\xi - \zeta|$
  - If $\lambda^\star = 0$, the sup is attained at $\zeta_i^\star = 0$ for all ξ_i, leading to $x^\star = 0 \longrightarrow \rho$ **too large, worst-case**
  - If $\lambda^\star \geq u$, the sup is attained at $\zeta_i^\star = \xi_i$ for each $\xi_i \longrightarrow$ **SAA problem** linear cost/function cancel out
  - $\lambda \in (0, u)$ cannot be optimal gradient either positive or negative
- WDRO leads to $x^\star_{WCR} = 0$ or $x^\star_{SAA}$ depending on ρ!

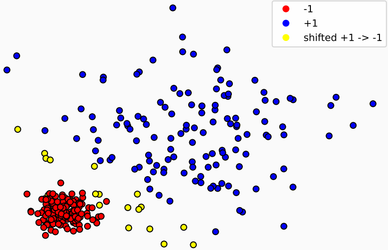
Example II – Logistic regression

- Standard classification problem
  - Labeled data $\xi_1, .., \xi_N$ of the form $\xi_i = (x_i, y_i) \in \mathbb{R}^d \times \{-1, +1\} = \Xi$
  - We minimize the loss $f_x(\xi = (x', y')) = \log(1 + \exp(-y'\langle x', x\rangle))$ by fitting separator $x \in \mathbb{R}^d$

$$\min_{x\in\mathbb{R}^d} \inf_{\lambda\geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^{N} \sup_{\zeta=(z,v)\in\Xi} \left\{\log(1+\exp(-y_i\langle x_i, x\rangle)) - \lambda\left(\|x_i - z\| + \kappa\mathbb{1}_{y_i\neq v}\right)\right\}$$

- We can solve Dual-WDRO by disciplined convex programming
  - for this, $c(\xi = (x, y), \zeta = (z, v)) = \|x - z\| + \kappa\mathbb{1}_{y\neq v}$ if $\kappa = +\infty$, (WDRO) is ERM regularized by $\rho\|x\|_*$

$$\begin{aligned}
\min_{x,\lambda,s} \quad & \lambda\rho + \frac{1}{N}\sum_{i=1}^{N} s_i \\
\text{s.t.} \quad & \log(1+\exp(-y_i\langle x_i, x\rangle)) \leq s_i \quad \forall i \\
& \log(1+\exp(y_i\langle x_i, x\rangle)) - \kappa\lambda \leq s_i \quad \forall i \\
& \|x\|_* \leq \lambda
\end{aligned}$$



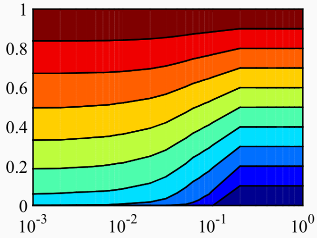
Example III – Portfolio selection

- Optimize a portfolio $x \in \{y \in \mathbb{R}^d_+ : \sum_{i=1}^d y[i] = 1\}$ over m assets subject to uncertain yearly returns
  - Return data $\xi_1, .., \xi_N$ in $\mathbb{R}^d = \Xi$
  - We minimize a risk-averse loss $f_x(\xi, \tau) = -\langle x, \xi \rangle + \eta\tau + \frac{\eta}{\alpha}\max(-\langle x, \xi\rangle - \tau; 0)$ with $\eta \geq 0$ is the risk aversion and $\alpha \in (0, 1]$ is the risk level $\rightsquigarrow$ risk $\mathbb{E}[-\langle x, \xi\rangle] + \eta \,\mathrm{CVaR}_\alpha[-\langle x, \xi\rangle]$

$$\min_{x \in \{\mathbb{R}^d_+ : \sum_{i=1}^d x[i]=1\}} \min_{\tau\in\mathbb{R}} \inf_{\lambda \geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^N \sup_{\zeta\in\Xi}\left\{-\langle x, \zeta\rangle + \eta\tau + \frac{\eta}{\alpha}\max(-\langle x, \zeta\rangle - \tau; 0) - \lambda\|\xi_i - \zeta\|\right\}$$

- We can again solve Dual-WDRO by disciplined convex programming for $c(\xi, \zeta) = \|\xi - \zeta\|$

$$\begin{aligned}
\min_{x,\tau,\lambda,s} \quad & \lambda\rho + \frac{1}{N}\sum_{i=1}^N s_i \\
\text{s.t.} \quad & \eta\tau - \langle x, \xi_i\rangle \leq s_i \quad \forall i \\
& \eta(1 - 1/\alpha)\tau - (1 + \eta/\alpha)\langle x, \xi_i\rangle \leq s_i \quad \forall i \\
& \|x\|_* \leq \lambda/\eta, \sum_{i=1}^d x[i] = 1, x \geq 0
\end{aligned}$$



Portfolio as a function of ρ Source: Esfahani & Kuhn, 2018

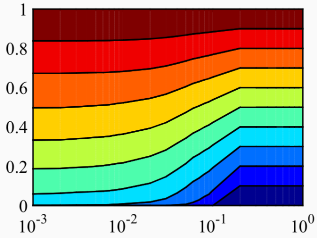
Example III – Portfolio selection

- Optimize a portfolio $x \in \{y \in \mathbb{R}^d_+ : \sum_{i=1}^d y[i] = 1\}$ over m assets subject to uncertain yearly returns
  - Return data $\xi_1, .., \xi_N$ in $\mathbb{R}^d = \Xi$
  - We minimize a risk-averse loss $f_x(\xi, \tau) = -\langle x, \xi \rangle + \eta\tau + \frac{\eta}{\alpha}\max(-\langle x, \xi\rangle - \tau; 0)$ with $\eta \geq 0$ is the risk aversion and $\alpha \in (0, 1]$ is the risk level $\rightsquigarrow$ risk $\mathbb{E}[-\langle x, \xi\rangle] + \eta \, \mathrm{CVaR}_\alpha[-\langle x, \xi\rangle]$

$$\min_{x \in \{\mathbb{R}^d_+ : \sum_{i=1}^d x[i]=1\}} \min_{\tau \in \mathbb{R}} \inf_{\lambda \geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^N \sup_{\zeta \in \Xi} \left\{ -\langle x, \zeta\rangle + \eta\tau + \frac{\eta}{\alpha}\max(-\langle x, \zeta\rangle - \tau; 0) - \lambda\|\xi_i - \zeta\| \right\}$$

- We can again solve Dual-WDRO by disciplined convex programming for $c(\xi, \zeta) = \|\xi - \zeta\|$
- Recovers that optimality of equally weighted portfolio under high ambiguity

- Esfahani and Kuhn. *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations.* Mathematical Programming, 2018.
- Pflug, Pichler, Wozabal. *The 1/N investment strategy is optimal under high model ambiguity.* J. Bank. Financ., 2012.
- Rockafellar and Uryasev. *Optimization of conditional value-at-risk.* J. Risk, 2000.



Portfolio as a function of ρ Source: Esfahani & Kuhn, 2018

Wasserstein Distributionally Robust Optimization

- **Statistical guarantees**

Statistical properties of WDRO: a first approach

- Let $\hat{\mathbf{P}}_N = \frac{1}{N}\sum_{i=1}^{N}\delta_{\xi_i}$ with $\xi_i \sim \mathbf{P}$ i.i.d.
  - We can see $\mathbf{P}$ as the *true* distribution encountered in practice
  - Take $c(\xi,\zeta) = \|\xi - \zeta\|^2$ and $\Xi \subset \mathbb{R}^d$ compact, convex, with nonempty interior
  - Concentration results for optimal transport $\mathbf{P}$ has finite moments

$$\mathbb{P}\left[W_2^2(\hat{\mathbf{P}}_N, \mathbf{P}) \leq \rho^2\right] \geq 1 - c_1 e^{-c_2 N \rho^d}$$

  - With probability at least $1-\delta$, taking $\rho \propto \frac{\log(1/\delta)}{N^{1/d}}$, for any f_x

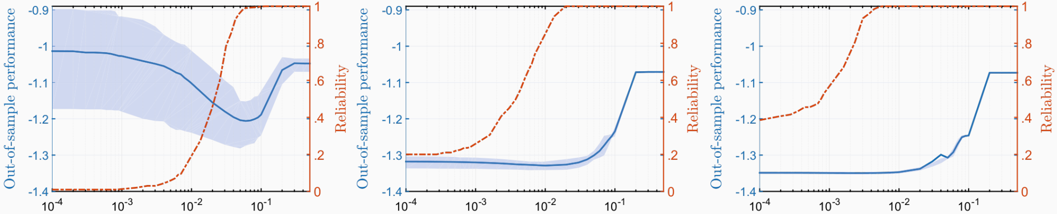
$$\textbf{unaccessible/target} \quad \mathbb{E}_{\xi\sim\mathbf{P}}[f_x(\xi)] \leq \widehat{\mathcal{R}}_\rho(f_x) = \sup_{\substack{\mathbf{Q}\in\mathcal{P}(\Xi)\\ W_2(\hat{\mathbf{P}}_N,\mathbf{Q})\leq\rho}} \mathbb{E}_{\xi\sim\mathbf{Q}}[f_x(\xi)] \quad \textbf{computable}$$

  - Overly pessimistic due to the curse of dimensionality N scales exponentially in d

- Fournier and Guillin. *On the rate of convergence in Wasserstein distance of the empirical measure.* Probability Theory and Related Fields, 2015
- Esfahani and Kuhn. *Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations.* Mathematical Programming, 2018.

- Sample 200 training datasets of size $N = \{30, 300, 3000\}$ from the same distribution
  - for each of them, solve WDRO to get optimal point $\hat{x}^\star$ and value $\widehat{\mathcal{R}}_\rho(f_{\hat{x}^\star})$
- **Reliability** = pc. of datasets s.t. the WDRO value is greater than the loss at the WDRO optimal point:

  estimated by taking $N = 30000$ **target** $\mathbb{E}_{\xi\sim\mathbf{P}}[f_{\hat{x}^\star}(\xi)] \leq \widehat{\mathcal{R}}_\rho(f_{\hat{x}^\star})$ **computed**



Out-of-sample performance $\mathbb{E}_{\xi\sim\mathbf{P}}\left[f_{\hat{x}^\star}(\xi)\right]$ and reliability as a function of ρ Source: Esfahani & Kuhn, 2018

- To get a fixed reliability, no need to scale $\frac{1}{N^{1/10}}$, $\frac{1}{\sqrt{N}}$ seems enough!

Statistical properties of WDRO: a finer proposition

- **Objective:** Get that with high probability $\mathbb{E}_{\xi\sim\mathbf{P}}[f_x(\xi)] \leq \widehat{\mathcal{R}}_\rho(f_x)$
- Why can we hope to do better than measure concentration ?
  - We do no *need* to bound the distance $W_2^2(\hat{\mathbf{P}}_N, \mathbf{P})$ whp.
  - Using the dual formulation, we can reformulate the target inequality as

$$\mathbb{E}_{\xi\sim\mathbf{P}}[f_x(\xi)] = \sup_{\substack{\mathbf{Q}\in\mathcal{P}(\Xi)\\ W_2(\mathbf{P},\mathbf{Q})\leq 0}} \mathbb{E}_{\xi\sim\mathbf{Q}}[f_x(\xi)]$$

$$= \inf_{\lambda\geq 0} \mathbb{E}_{\xi\sim\mathbf{P}}\left[\sup_{\zeta\in\Xi}\left\{f_x(\zeta) - \lambda\|\xi-\zeta\|^2\right\}\right] \leq \inf_{\lambda\geq 0} \lambda\rho + \mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\sup_{\zeta\in\Xi}\left\{f_x(\zeta) - \lambda\|\xi-\zeta\|^2\right\}\right]$$

$$= \widehat{\mathcal{R}}_\rho(f_x)$$

  - *Rather*, we may show that whp **uniformly in f_x** (ok...) **and in λ** (less cool)

$$\frac{\mathbb{E}_{\xi\sim\mathbf{P}}\left[\sup_{\zeta\in\Xi}\left\{f_x(\zeta) - \lambda\|\xi-\zeta\|^2\right\}\right]}{\lambda} \leq \frac{\mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\sup_{\zeta\in\Xi}\left\{f_x(\zeta) - \lambda\|\xi-\zeta\|^2\right\}\right]}{\lambda} + \rho$$

  - The concentration error is directly related to the radius ρ hopefully with a good dependency in N...

Concentration

- For $\delta \in (0,1)$ and some $0 < \underline{\lambda} < +\infty$, we have with probability at least $1-\delta/2$ that

$$\sup_{(f_x,\lambda)\in\mathcal{F}\times[\underline{\lambda},+\infty)} \left\{ \frac{\mathbb{E}_{\xi\sim\mathbf{P}}\left[\sup_{\zeta\in\Xi}\left\{f_x(\zeta)-\lambda\|\xi-\zeta\|^2\right\}\right] - \mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\sup_{\zeta\in\Xi}\left\{f_x(\zeta)-\lambda\|\xi-\zeta\|^2\right\}\right]}{\lambda} \right\}$$

$$\leq \frac{117}{\sqrt{N}\underline{\lambda}}\left(\mathcal{I}(\mathcal{F}) + \mathrm{Cst}\left(1+\sqrt{\log\frac{1}{\delta}}\right)\right) := \rho_N$$

  - **Error $\rightsquigarrow$ minimal radius for concentration in $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$** no curse of dimension
  - The complexity of the class of functions appears as one can expect
  - **Lower bound on the dual variable, $\underline{\lambda}$, needed** we have to show it
- Proof relies on standard concentration results + sup Lipschitz and bounded

◇ Boucheron, Lugosi, and Massart. *Concentration Inequalities. A Nonasymptotic Theory of Independence.* Oxford University Press, 2013.

Lower-bounding the dual variable

- Crux of the proof for getting the concentration result: **if** $\lambda > \underline{\lambda}$, we have with probability $1 - \delta/2$

$$\mathbb{E}_{\xi \sim \mathbf{P}}\left[f_x(\xi)\right] \leq \widehat{\mathcal{R}}_\rho(f_x)$$

whenever ρ is bigger than $\rho_N = \frac{117}{\sqrt{N}\underline{\lambda}}\left(\mathcal{I}(\mathcal{F}) + \mathrm{Cst}\left(1 + \sqrt{\log \frac{1}{\delta}}\right)\right).$

- Careful analysis of the dual function: It's all a matter of compromises
  - If ρ is small, the constraint is stringent, λ is big
    BUT ρ has to be also greater than ρ_N
  - If ρ is bigger, we have more margin for error
    BUT the constraint has to be sufficiently active so that λ does not vanish
  - The lower bound depends on $\hat{\mathbf{P}}_N$ natively which is not nice
    SO we also need to concentrate the opposite quantity to get back to $\mathbf{P}$
  - AND the minimal λ depends on ρ...

Theorem (Azizian, I., Malick'23 – informal)
There is a critical radius ρ_c depending only on $\mathcal{F}$ and $\mathbf{P}$ such that for any $\delta \in (0,1)$ and $N \geq 1$, if

$$\mathcal{O}\left(\sqrt{\frac{1+\log 1/\delta}{N}}\right) \leq \rho \leq \frac{\rho_c}{2} - \mathcal{O}\left(\sqrt{\frac{1+\log 1/\delta}{N}}\right)$$

then, there is $\rho_N = \mathcal{O}\left(\sqrt{\frac{1+\log 1/\delta}{N}}\right)$ such that, with probability $1-\delta$, $\forall f_x \in \mathcal{F}$,

$$\mathbb{E}_{\xi\sim\mathbf{Q}}\left[f_x(\xi)\right] \leq \widehat{\mathcal{R}}_\rho(f_x) \qquad \textit{for all } \mathbf{Q} \textit{ such that } W_2(\mathbf{P},\mathbf{Q}) \leq \sqrt{\rho(\rho-\rho_N)},$$

In particular, with probability $1-\delta$, we have $\forall f_x \in \mathcal{F}$,

$$\mathbb{E}_{\xi\sim\mathbf{P}}\left[f_x(\xi)\right] \leq \widehat{\mathcal{R}}_\rho(f_x).$$

- $\rho_c = \sqrt{\inf_{f_x\in\mathcal{F}} \mathbb{E}_{\xi\sim\mathbf{P}}\left[\frac{1}{2}d^2(\xi, \arg\max f_x)\right]}$ is the maximal radius before falling back to worst case robustness

A word on assumptions and literature

- Assumptions
  - Sample space Ξ compact convex + supp $\mathbf{P}$ strictly included in Ξ with some margin
  - All functions f_x are twice differentiable + bounded/smooth regularized or non-convex models are ok
  - Decrease condition around maximizers of f_x uniformly + non-vanishing gradients ok for most linear models
- Literature: Bridging the gap between several results on concentration for WDRO with $\rho \propto 1/\sqrt{N}$
  - With error terms
  - Asymptotic
  - Experimental

◇ An and Gao. *Generalization Bounds for (Wasserstein) Robust Optimization.* NeurIPS, 2021.
◇ Blanchet, Murthy, and Si. *Confidence regions in wasserstein distributionally robust estimation.* Biometrika, 2022.

Conclusion on statistical guarantees

- WDRO models control the true risk with high probability
  - Radius ρ should be taken proportional to $1/\sqrt{N}$
  - Uniform in the model f_x we still have to optimize it!
- What about tightness?
  - Under the same assumptions whp.

$$\mathcal{R}_{\sqrt{\rho(\rho-\rho_N)}}(f_x) \leq \widehat{\mathcal{R}}_\rho(f_x) \leq \mathcal{R}_{\sqrt{\rho(\rho+\rho_N)}}(f_x)$$

with $\mathcal{R}_\rho(f_x)$ the (regularized) WDRO risk rooted at $\mathbf{P}$

Wasserstein Distributionally Robust Optimization

◇ **Approximation**

Entropic regularization

- We wish to get rid of the **linearity** of the problem
  - We draw inspiration from **regularization in optimal transport**

WDRO

$$\sup_{\substack{\mathbf{Q}\in\mathcal{P}(\Xi)\\ W_c(\hat{\mathbf{P}}_N,\mathbf{Q})\leq\rho}} \mathbb{E}_{\xi\sim\mathbf{Q}}[f_x(\xi)]$$

$$= \inf_{\lambda\geq 0} \lambda\rho + \mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\sup_{\zeta\in\Xi}\{f_x(\zeta)-\lambda c(\xi,\zeta)\}\right]$$

$$:= \widehat{\mathcal{R}}_\rho(f_x)$$

- Peyré, Cuturi. *Computational Optimal Transport.* Foundation and Trends in Machine Learning, 2019.
- Wang, Gao, and Xie. *Sinkhorn Distributionally Robust Optimization.* ArXiv 2109.11926, 2021.
- Azizian, I., Malick. *Regularization for Wasserstein Distributionally Robust Optimization.* ESAIM:COCV, 2022.
- Piat, Fadili, Jurie, da Veiga. *Regularized Robust Optimization with Application to Robust Learning.* preprint, 2022.

Entropic regularization

- We wish to get rid of the **linearity** of the problem
  - We draw inspiration from **regularization in optimal transport**

WDRO

$$\sup_{\substack{\pi \in \mathcal{M}(\Xi\times\Xi) \\ \pi_1 = \hat{\mathbf{P}}_N,\ \langle \pi, c\rangle \leq \rho}} \langle \pi, \tilde{f}_x \rangle$$

$$= \inf_{\lambda \geq 0} \lambda\rho + \mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\sup_{\zeta\in\Xi} \{f_x(\zeta) - \lambda c(\xi,\zeta)\}\right]$$

$$:= \widehat{\mathcal{R}}_\rho(f_x)$$

Regularized WDRO

$$\sup_{\substack{\pi \in \mathcal{M}(\Xi\times\Xi) \\ \pi_1 = \hat{\mathbf{P}}_N,\ \langle \pi, c\rangle \leq \rho}} \langle \pi, \tilde{f}_x \rangle - \varepsilon\,\mathrm{KL}(\pi\,|\pi_0)$$

$$:= \widehat{\mathcal{R}}_\rho^\varepsilon(f_x)$$

  - π must be absolutely continuous wrt. to the *chosen* π_0 and $(\pi_0)_1 = \hat{\mathbf{P}}_N$

- Peyré, Cuturi. *Computational Optimal Transport.* Foundation and Trends in Machine Learning, 2019.
- Wang, Gao, and Xie. *Sinkhorn Distributionally Robust Optimization.* ArXiv 2109.11926, 2021.
- Azizian, I., Malick. *Regularization for Wasserstein Distributionally Robust Optimization.* ESAIM:COCV, 2022.
- Piat, Fadili, Jurie, da Veiga. *Regularized Robust Optimization with Application to Robust Learning.* preprint, 2022.

Entropic regularization

- We wish to get rid of the **linearity** of the problem
  - We draw inspiration from **regularization in optimal transport**

WDRO

$$\sup_{\substack{\pi\in\mathcal{M}(\Xi\times\Xi)\\ \pi_1=\hat{\mathbf{P}}_N,\ \langle\pi,c\rangle\le\rho}} \langle\pi,\tilde{f}_x\rangle$$

$$= \inf_{\lambda\ge 0} \lambda\rho + \mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\sup_{\zeta\in\Xi}\{f_x(\zeta)-\lambda c(\xi,\zeta)\}\right]$$

$$:= \widehat{\mathcal{R}}_\rho(f_x)$$

Regularized WDRO

$$\sup_{\substack{\pi\in\mathcal{M}(\Xi\times\Xi)\\ \pi_1=\hat{\mathbf{P}}_N,\ \langle\pi,c\rangle\le\rho}} \langle\pi,\tilde{f}_x\rangle - \varepsilon\,\mathrm{KL}(\pi\,|\pi_0)$$

$$= \inf_{\lambda\ge 0} \lambda\rho + \varepsilon\,\mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi)}\left[e^{\frac{f_x(\zeta)-\lambda c(\xi,\zeta)}{\varepsilon}}\right]\right)\right]$$

$$:= \widehat{\mathcal{R}}^{\varepsilon}_\rho(f_x)$$

  - π must be absolutely continuous wrt. to the *chosen* π_0 and $(\pi_0)_1 = \hat{\mathbf{P}}_N$
  - Lagrangian then Fenchel duality in the space of finite signed **measures** on a compact space
  - The sup is transformed into a log-integral-exp

- Peyré, Cuturi. *Computational Optimal Transport*. Foundation and Trends in Machine Learning, 2019.
- Wang, Gao, and Xie. *Sinkhorn Distributionally Robust Optimization*. ArXiv 2109.11926, 2021.
- Azizian, I., Malick. *Regularization for Wasserstein Distributionally Robust Optimization*. ESAIM:COCV, 2022.
- Piat, Fadili, Jurie, da Veiga. *Regularized Robust Optimization with Application to Robust Learning*. preprint, 2022.

Approximation error

- The reference coupling π_0 is a kind of prior
  - In optimal transport, entropic regularization with $\mathrm{KL}(\pi \mid \mathbf{P} \otimes \mathbf{Q})$ π_0 is the product of marginals
  - In WDRO, the second marginal is **not fixed** but optimized to get our adversarial distribution
  - We choose $\pi_0(\mathrm{d}\xi, \mathrm{d}\zeta) \propto \hat{\mathbf{P}}_N(\mathrm{d}\xi) e^{-\frac{\|\xi-\zeta\|^p}{2^{p-1}\sigma}} \mathbb{1}_{\zeta \in \Xi}\, \mathrm{d}\zeta$

$$\widehat{\mathcal{R}}_\rho(f_x) = \inf_{\lambda \geq 0} \lambda\rho + \mathbb{E}_{\xi \sim \hat{\mathbf{P}}_N}\left[\sup_{\zeta \in \Xi} \{f_x(\zeta) - \lambda\|\xi - \zeta\|^p\}\right] \tag{WDRO}$$

$$\widehat{\mathcal{R}}^\varepsilon_\rho(f_x) = \inf_{\lambda \geq 0} \lambda\rho + \varepsilon\, \mathbb{E}_{\xi \sim \hat{\mathbf{P}}_N}\left[\log\left(\mathbb{E}_{\zeta \sim \pi_0(\cdot|\xi)}\left[e^{\frac{f_x(\zeta) - \lambda\|\xi-\zeta\|^p}{\varepsilon}}\right]\right)\right] \tag{ε-WDRO}$$

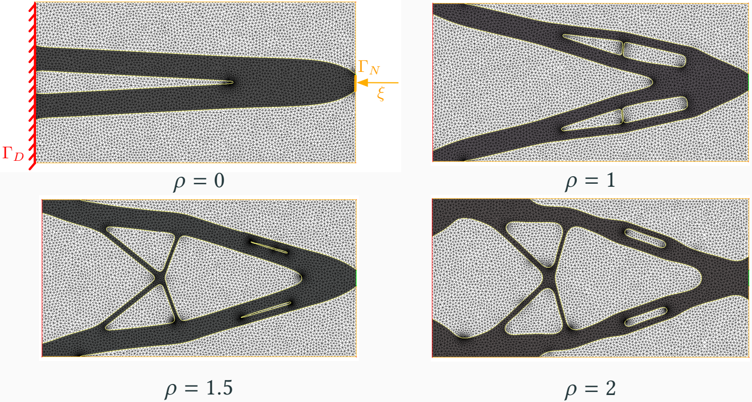
Theorem (Azizian, I., Malick'22)
If $\Xi \subset \mathbb{R}^d$ is compact, convex, with nonempty interior and f_x is Lipschitz continuous, then as ε goes to 0

$$0 \leq \widehat{\mathcal{R}}_\rho(f_x) - \widehat{\mathcal{R}}^\varepsilon_\rho(f_x) \leq \mathcal{O}\left(\varepsilon d \log\left(\frac{1}{\varepsilon}\right)\right)$$

◇ Genevay, Chizat, Bach, Cuturi, and Peyré. *Sample complexity of sinkhorn divergences.* AIStats, 2019.

Example IV – A problem that has no tractable WDRO formulation

- Optimization of a cantilever beam minimization of the compliance under a volume constraint
  - Uncertainty lies in the load ξ applied around the vector $(-1, 0)$
- Entropic WDRO formulation over a finite element solver Expectation approx. by a 10 Gaussian samples



◇ Dapogny, I., Meda, Thibert. *Entropy-regularized Wasserstein distributionally robust shape and topology optimization.* Structural and Multidisciplinary Optimization, ArXiv 2209.01500, 2022.

What about generalization?

- Thanks to our duality results, we can use the same kind of technique
  - **Classical**

$$\frac{\mathbb{E}_{\xi\sim\mathbf{P}}\left[\sup_{\zeta\in\Xi}\left\{f_x(\zeta)-\lambda\|\xi-\zeta\|^2\right\}\right]}{\lambda}\leq\frac{\mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\sup_{\zeta\in\Xi}\left\{f_x(\zeta)-\lambda\|\xi-\zeta\|^2\right\}\right]}{\lambda}+\rho$$

  - **Regularized** with $-\varepsilon\,\mathrm{KL}(\pi\,|\pi_0)$ and $\pi_0(\mathrm{d}\xi,\mathrm{d}\zeta)\propto\hat{\mathbf{P}}_N(\mathrm{d}\xi)e^{-\frac{\|\xi-\zeta\|^2}{2\sigma}}\mathbb{1}_{\zeta\in\Xi}\,\mathrm{d}\zeta$

$$\frac{\varepsilon\,\mathbb{E}_{\xi\sim\mathbf{P}}\left[\log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi)}\left[e^{\frac{f_x(\zeta)-\lambda\|\xi-\zeta\|^2}{\varepsilon}}\right]\right)\right]}{\lambda}\leq\frac{\varepsilon\,\mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi)}\left[e^{\frac{f_x(\zeta)-\lambda\|\xi-\zeta\|^2}{\varepsilon}}\right]\right)\right]}{\lambda}+\rho$$

- Same proof layout but quite different derivations
  - The additional parameters ε and σ should be taken **proportional to ρ** to get close to the true risk *at the same time* it naturally appears in the proofs

Concentration result for regularized WDRO

Theorem (Azizian, I., Malick'23 – informal)
For $\sigma = \sigma_0 \rho$ with $\sigma_0 > 0$, $\varepsilon = \varepsilon_0 \rho$ with $\varepsilon_0 > 0$ such that $\varepsilon_0 / \sigma_0^2$ is small enough depending on $\mathcal{F}$, $\mathbf{P}$, Ξ, there is an explicit constant ρ_c depending only on $\mathcal{F}$, $\mathbf{P}$ and Ξ such that for all $\delta \in (0,1)$ and $N \geq 1$, if

$$\mathcal{O}\left(\sqrt{\frac{1+\log 1/\delta}{N}}\right) \leq \rho \leq \frac{\rho_c}{2} - \mathcal{O}\left(\frac{1}{\sqrt{N}}\right), \quad \text{and} \quad \rho_c \geq \mathcal{O}\left(\frac{1}{N^{1/6}} + \left(\frac{1+\log 1/\delta}{N}\right)^{1/4}\right),$$

then, there are $\tau = \mathcal{O}(\varepsilon\rho)$ and $\rho_N = \mathcal{O}\left(\sqrt{\frac{1+\log 1/\delta}{N}}\right)$ such that, with probability at least $1-\delta$, $\forall f_x \in \mathcal{F}$,

$$\mathbb{E}_{\xi \sim \mathbf{Q}}\left[f_x(\xi)\right] \leq \widehat{\mathcal{R}}_\rho^\varepsilon(f_x) \qquad \text{for all } \mathbf{Q} \text{ such that } W_{2,\tau}(\mathbf{P}, \mathbf{Q}) \leq \sqrt{\rho(\rho - \rho_N)}$$

Furthermore, when σ_0 and σ are small enough depending on $\mathbf{P}$ and Ξ, with probability $1-\delta$, $\forall f_x \in \mathcal{F}$,

$$\mathbb{E}_{\xi \sim \mathbf{P}} \mathbb{E}_{\zeta \sim \pi_0(\cdot|\xi)}\left[f_x(\zeta)\right] \leq \widehat{\mathcal{R}}_\rho^\varepsilon(f_x).$$

- Not exactly an upper bound on the true risk on $\mathbf{P}$ but rather the risk for smoothed $\mathbf{P} * \pi_0(\cdot|\xi)$
- Robust wrt. $W_{2,\tau}(\mathbf{P}, \mathbf{Q}) := \sqrt{\inf\left\{\mathbb{E}_\pi\left[\frac{1}{2}\|\xi - \zeta\|^2\right] + \tau\, \mathrm{KL}(\pi|\pi_0) : \pi \in \mathcal{P}(\Xi \times \Xi), \pi_1 = \mathbf{P}, \pi_2 = \mathbf{Q}\right\}}$

Conclusion on approximation

- The WDRO problem $\widehat{\mathcal{R}}_\rho(f_x)$ can be controllably approximated by

$$\widehat{\mathcal{R}}_\rho^\varepsilon(f_x) = \sup_{\substack{\pi\in\mathcal{M}(\Xi\times\Xi)\\ \pi_1=\hat{\mathbf{P}}_N,\ \langle\pi,c\rangle\leq\rho}} \langle\pi,\tilde{f}_x\rangle \ -\varepsilon\,\mathrm{KL}(\pi\,|\pi_0) = \inf_{\lambda\geq 0} \lambda\rho + \varepsilon\,\mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi)}\left[e^{\frac{f_x(\zeta)-\lambda\|\xi-\zeta\|^p}{\varepsilon}}\right]\right)\right]$$

  - Differentiable and more tractable problem as soon as the inner integral can be evaluated
- This is not exactly a Sinkhorn distance
  - We can regularize in the objective and/or in the constraints
  - We cannot symmetrize to get an actual distance
- Worst case probability measures from optimal dual value $\lambda^\star$

$$\propto \sum_{i=1}^N e^{\frac{f_x(\zeta)-\lambda^\star\|\xi_i-\zeta\|^p}{\varepsilon}}\, e^{-\frac{\|\xi_i-\zeta\|^p}{2^{p-1}\sigma}}\,\mathbb{1}_{\zeta\in\Xi}\,\mathrm{d}\zeta$$

- Concentration is very similar for the regularized version
  - Insight on the choice of $\varepsilon\propto\rho$ same thing for σ
  - Thanks to regularization, we get rid of the need to control the behavior near maximizers

Wasserstein Distributionally Robust Optimization

- ◊ **Optimization**

Solving generic WDRO problems

- Leverage the entropic regularization

$$\min_{x\in\mathcal{X}} \inf_{\lambda\geq 0} \lambda\rho + \varepsilon \frac{1}{N}\sum_{i=1}^{N}\left[\log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi_i)}\left[e^{\frac{f_x(\zeta)-\lambda\|\xi_i-\zeta\|^2}{\varepsilon}}\right]\right)\right]$$

  - Gradients in x and λ are available

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi_i)}\nabla_x f_x(\zeta)e^{\frac{f_x(\zeta)-\lambda\|\xi_i-\zeta\|^2}{\varepsilon}}}{\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi_i)}e^{\frac{f_x(\zeta)-\lambda\|\xi_i-\zeta\|^2}{\varepsilon}}}\right] \text{ and } \rho - \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi_i)}\|\xi_i-\zeta\|^2 e^{\frac{f_x(\zeta)-\lambda\|\xi_i-\zeta\|^2}{\varepsilon}}}{\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi_i)}e^{\frac{f_x(\zeta)-\lambda\|\xi_i-\zeta\|^2}{\varepsilon}}}\right]$$

- **Crude approach:** sample some points from $\pi_0(\cdot|\xi_i) \propto e^{\frac{\|\xi_i-\zeta\|^2}{2\sigma}} \mathbb{1}_{\zeta\in\Xi}$ and minimize the sampled loss
  - This is a biased approximation with poor performance in practice except for $d = 1$
- **Better approach:** sample the expectation at each iteration by (Metropolis-adjusted) Langevin
  - “Robustifies” but unstable behavior of λ
- **Implemented approach:** additionally use importance sampling towards $\nabla_{\xi_i} f_x(\xi_i)$
  - Much more stable, when initialized with the ERM solution

Talking code

- Python package coming soon – Two modes:
  - *à la* scikit-learn

```
from sklearn.linear_model import LogisticRegression # scikit-learn's standard version
from skwdro.linear_models import LogisticRegression as WDROLogisticRegression # WDRO version
```

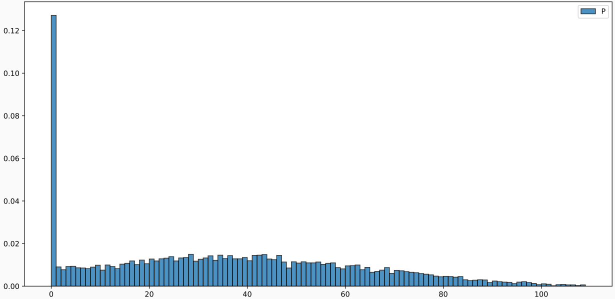
  - *à la* pytorch

```
from typing import Optional
from abc import abstractclassmethod, abstractproperty

import torch as pt
import torch.nn as nn

from skwdro.base.samplers.torch.base_samplers import BaseSampler

class Loss(nn.Module):
    """ Base class for loss functions """
    _sampler: BaseSampler
    def __init__(self, sampler: BaseSampler):
        super(Loss, self).__init__()
        self._sampler = sampler

    def value(self, xi: pt.Tensor, xi_labels: Optional[pt.Tensor]):
        """
        Perform forward pass.
        """
        raise NotImplementedError("Please Implement this method")
```

Back to example I – the NewsVendor problem

- A NewsVendor has to maximize its gain $-f_x(\xi) = -kx + u\min(x,\xi)$ by optimizing the number $x \in \mathbb{R}_+$ of newspaper bought, facing the uncertain demand of tomorrow
  - His buying price is $k = 5$ and his retail price is $u = 7$ $\rho = 2, \varepsilon = 0.1$
  - $N = 20$ i.i.d. samples from $\mathbf{P}$
- Samples distribution $\mathbf{P}$: "good day" $\mathcal{N}(50,5)$ w/ prob. 0.5, "bad day" $\mathcal{N}(20,5)$ w/ prob. 0.5, truncated at 0

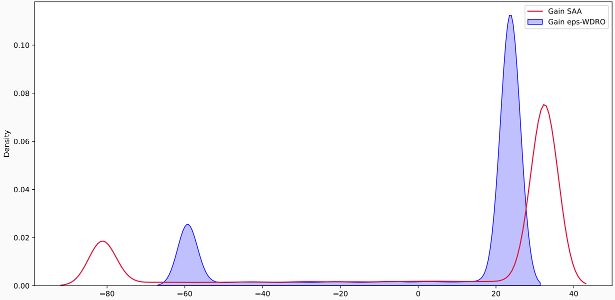
	SAA	WDRO	ε-WDRO
$x^\star$	16	0	12
Empirical loss	11.10	0.00	9.99



Back to example I – the NewsVendor problem

- A NewsVendor has to maximize its gain $-f_x(\xi) = -kx + u\min(x,\xi)$ by optimizing the number $x \in \mathbb{R}_+$ of newspaper bought, facing the uncertain demand of tomorrow
  - His buying price is $k = 5$ and his retail price is $u = 7$ $\rho = 2, \varepsilon = 0.1$
  - $N = 20$ i.i.d. samples from $\mathbf{P}$
- Samples distribution $\mathbf{P}$: "good day" $\mathcal{N}(50, 5)$ w/ prob. 0.5, "bad day" $\mathcal{N}(20, 5)$ w/ prob. 0.5, truncated at 0

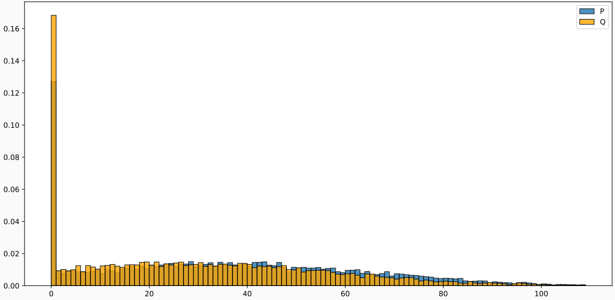
	SAA	WDRO	ε-WDRO
$x^\star$	16	0	12
Empirical loss	11.10	0.00	9.99
Actual gain on $\mathbf{P}$	11.04	0.00	9.78



Back to example I – the NewsVendor problem

- A NewsVendor has to maximize its gain $-f_x(\xi) = -kx + u\min(x, \xi)$ by optimizing the number $x \in \mathbb{R}_+$ of newspaper bought, facing the uncertain demand of tomorrow
  - His buying price is $k = 5$ and his retail price is $u = 7$ $\rho = 2, \varepsilon = 0.1$
  - $N = 20$ i.i.d. samples from $\mathbf{P}$
- Samples distribution $\mathbf{P}$: "good day" $\mathcal{N}(50, 5)$ w/ prob. 0.5, "bad day" $\mathcal{N}(20, 5)$ w/ prob. 0.5, truncated at 0
- Shifted distribution $\mathbf{Q}$: "good day" $\mathcal{N}(50, 5)$ w/ prob. 0.3, "bad day" $\mathcal{N}(20, 5)$ w/ prob. 0.7, truncated at 0

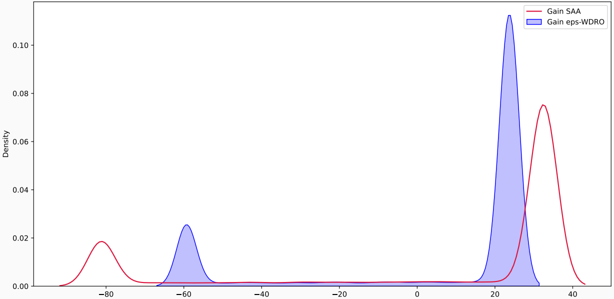
	SAA	WDRO	ε-WDRO
$x^\star$	16	0	12
Empirical loss	11.10	0.00	9.99
Actual gain on $\mathbf{P}$	11.04	0.00	9.78



Back to example I – the NewsVendor problem

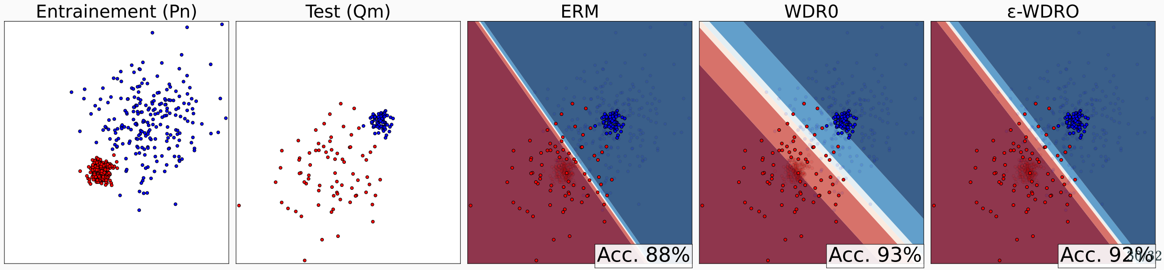
- A NewsVendor has to maximize its gain $-f_x(\xi) = -kx + u\min(x, \xi)$ by optimizing the number $x \in \mathbb{R}_+$ of newspaper bought, facing the uncertain demand of tomorrow
  - His buying price is $k = 5$ and his retail price is $u = 7$ $\rho = 2, \varepsilon = 0.1$
  - $N = 20$ i.i.d. samples from **P**
- Samples distribution **P**: "good day" $\mathcal{N}(50, 5)$ w/ prob. 0.5, "bad day" $\mathcal{N}(20, 5)$ w/ prob. 0.5, truncated at 0
- Shifted distribution **Q**: "good day" $\mathcal{N}(50, 5)$ w/ prob. 0.3, "bad day" $\mathcal{N}(20, 5)$ w/ prob. 0.7, truncated at 0

	SAA	WDRO	ε-WDRO
$x^\star$	16	0	12
Empirical loss	11.10	0.00	9.99
Actual gain on **P**	11.04	0.00	9.78
Actual gain on **Q**	5.65	0.00	6.08



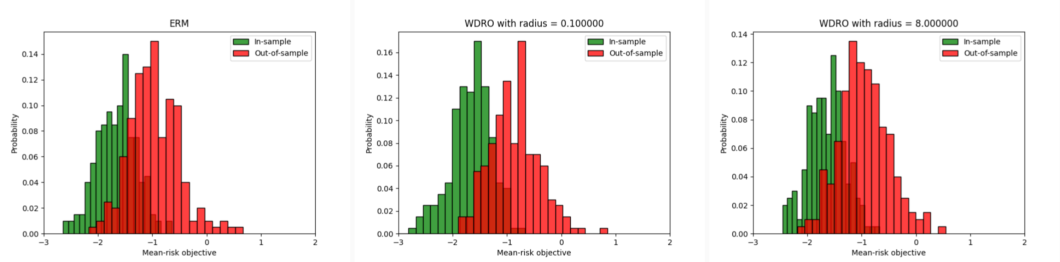
Back to example II – Logistic regression

- Regularization offers new possibilities:
  - Different transport costs squared norm, exotic ones
  - Regularization l1, l2, anything not data-driven
  - Scaling to larger datasets gradient-based methods instead of DCP
- Regularized WDRO as new robustness model:
  - ε is not necessarily small $\max(1e^{-3}, \rho/10)$
  - Absolutely continuous true distribution prior linked to transport cost



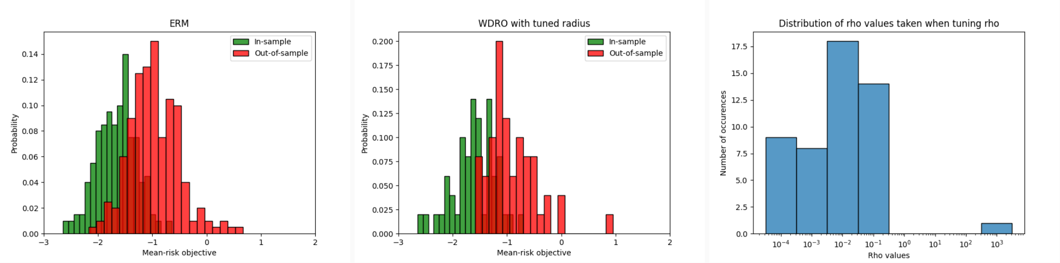
Back to example III – Portfolio selection

- 10 assets, $N = 30$, 200 simulations
- Choice of the radius ρ
  - Cross validation inherited as a scikit-learn estimator
  - By statistically testing that the test distribution is encompassed



Back to example III – Portfolio selection

- 10 assets, $N = 30$, 200 simulations
- Choice of the radius ρ
  - Cross validation inherited as a scikit-learn estimator
  - By statistically testing that the test distribution is encompassed



Conclusion on optimization

- Why optimizing correctly matters?
  - Being accurate in λ enables to get a bound on the true risk
  - Instrumental to get the worst case distributions
- Toolbox underway!
  - Based on solving the dual regularized problem
  - AdamW using importance sampling for approximating the expectation
  - Default values for the radius, regularization, etc. from statistical study
- Currently undergoing testing on optimization and generalization
  - **email me if you're interested in a β-version**
- Paves the way to a widespread use of WDRO
  - Large class of objectives and costs not necessarily smooth
  - Cross validation of parameters

Conclusion

Hokusai

Fine Wind, Clear Morning (Gaifū kaisei) in Thirty-six Views of Mount Fuji (1830-1832)

Closing words

- Machine Learning models perform well but are they reliable?
  - Distributionally robust optimization provides an appealing framework to address this question
  - Interplay between statistics and optimization
- Wasserstein distributionally robust models are in!
  - Generalization and robustness guarantees
  - Widely implementable thanks to regularization
- Exciting perspectives: automated radius tuning, practical applications, robust feature selection, etc.

Azizian, I., Malick: *Regularization for Wasserstein Distributionally Robust Optimization*, arXiv 2205.08826, ESAIM: Control, Optimisation, and Calculus of Variations, 2023.

Azizian, I., Malick: *Exact Generalization Guarantees for (Regularized) Wasserstein Distributionally Robust Models*, arXiv 2305.17076, NeurIPS, 2023.

Dapogny, I., Meda, Thibert. *Entropy-regularized Wasserstein distributionally robust shape and topology optimization.* ArXiv 2209.01500, Structural and Multidisciplinary Optimization, 2022.

Thank you! – www.iutzeler.org